

SMiP Core Course:  
Foundations of Statistical Modeling II

# Multinomial Processing Tree Modeling: Basic Methods and Recent Advances

Edgar Erdfelder (Mannheim) & Daniel W. Heck (Marburg)



STATISTICAL MODELING in PSYCHOLOGY  
FREIBURG HEIDELBERG LANDAU MANNHEIM TÜBINGEN



# Multinomial Processing Tree (MPT) Modeling, Part 1: Basics

Edgar Erdfelder and Daniel W. Heck

University of Mannheim & Phillips University Marburg

# 1) Basics

- 1.1) Introduction to standard MPT models
- 1.2) Examples
- 1.3) Model development
- 1.4) Formal model structure
- 1.5) Identifiability
- 1.6) Parameter estimation
- 1.7) Model assessment
- 1.8) Selected literature

# 1.1) Introduction to standard MPT models

- **Required type of data:**
- Standard multinomial models are tailored to discrete (i.e., categorical) data.
- Psychological data are typically discrete in nature (e.g., yes/no responses, correct/incorrect judgments, ratings, choices, ...).
- If not, they can often be transformed into discrete data
  - Response times: Categorization into bins
  - Numerical judgments: Rank-orders of judgments
- Hence, many psychological paradigms generate frequency data that are appropriate for MPT modeling.

# 1.1) Introduction to standard MPT models

- **Distributional assumptions:**
- Standard MPT models assume that observations are sampled independently from
  - one multinomial distribution (simple multinomial model)
  - several multinomial distributions (joint multinomial model)
- This includes simple and joint binomial models as special cases.
- The frequency data structure can be univariate or multivariate.

# MPT models ...

- ... provide explanations of observed frequency data in terms of basic parameters with clear-cut psychological interpretations;
- ... these parameters represent probabilities of latent psychological processes (or latent psychological states) underlying human behavior;
- ... in other words, these models measure the contributions of different psychological processes to frequencies of observable behaviors.
- In this sense, multinomial models allow for a “measurement of cognitive processes” (Riefer & Batchelder, 1988)

## 1.2) Examples

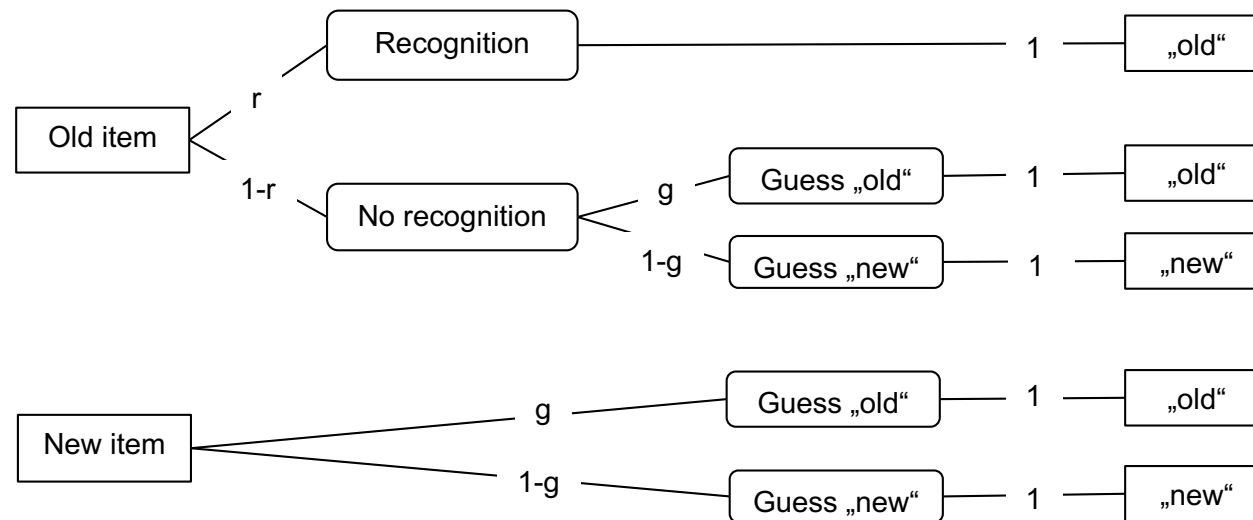
A very simple example:

- *Paradigm:*
  - Yes-No recognition test
- *Two Conditions:*
  - Old Items
  - New Items
- *Categorical (dichotomous) dependent variable:*
  - „Old“ vs. „New“ Judgment

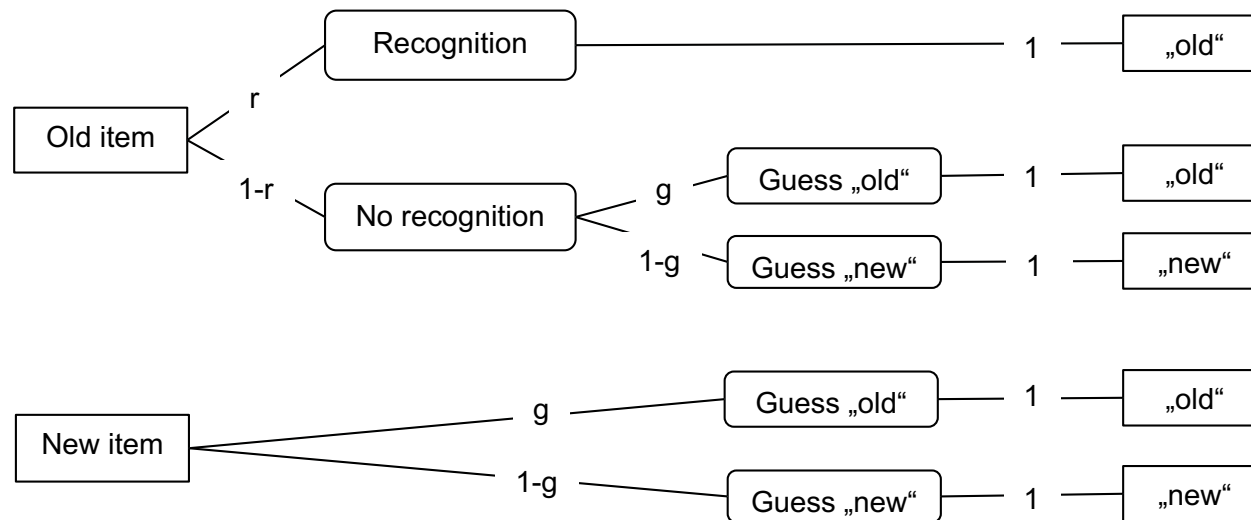
# A) Separating recognition and guessing: The One-High Threshold Model (Blackwell, 1963)



# A) Separating recognition and guessing: The One-High Threshold Model (Blackwell, 1963)



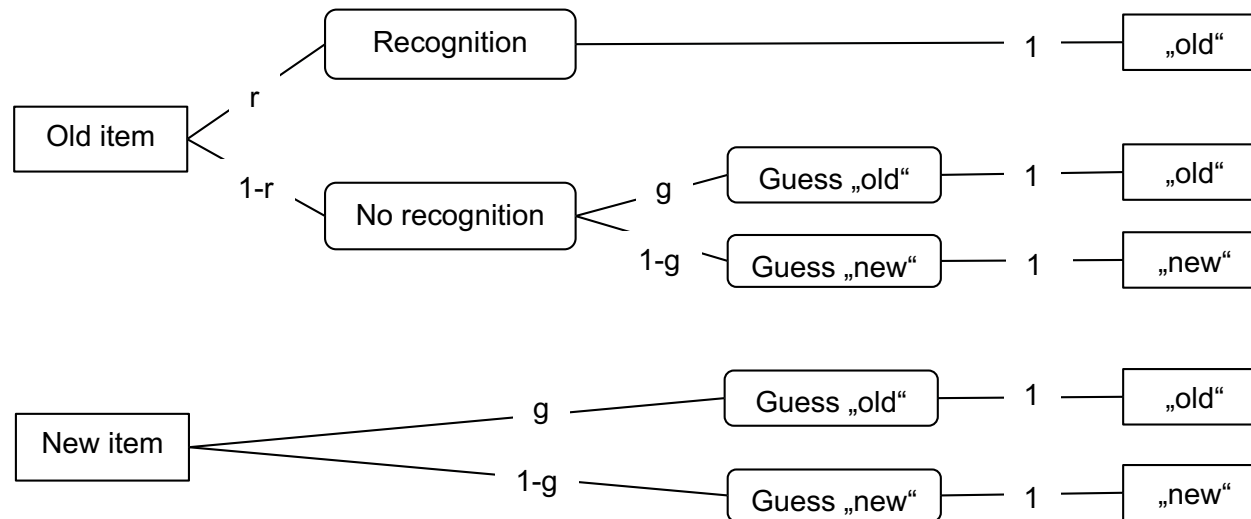
# A) Separating recognition and guessing: The One-High Threshold Model (Blackwell, 1963)



*Model equations:*

$$p(\text{„old“} \mid \text{old item}) = r + (1-r) \cdot g$$

# A) Separating recognition and guessing: The One-High Threshold Model (Blackwell, 1963)



*Model equations:*

$$p(\text{„old“} \mid \text{old item}) = r + (1-r) \cdot g$$

$$p(\text{„old“} \mid \text{new item}) = g$$

## B) Separating storage and retrieval in long-term memory

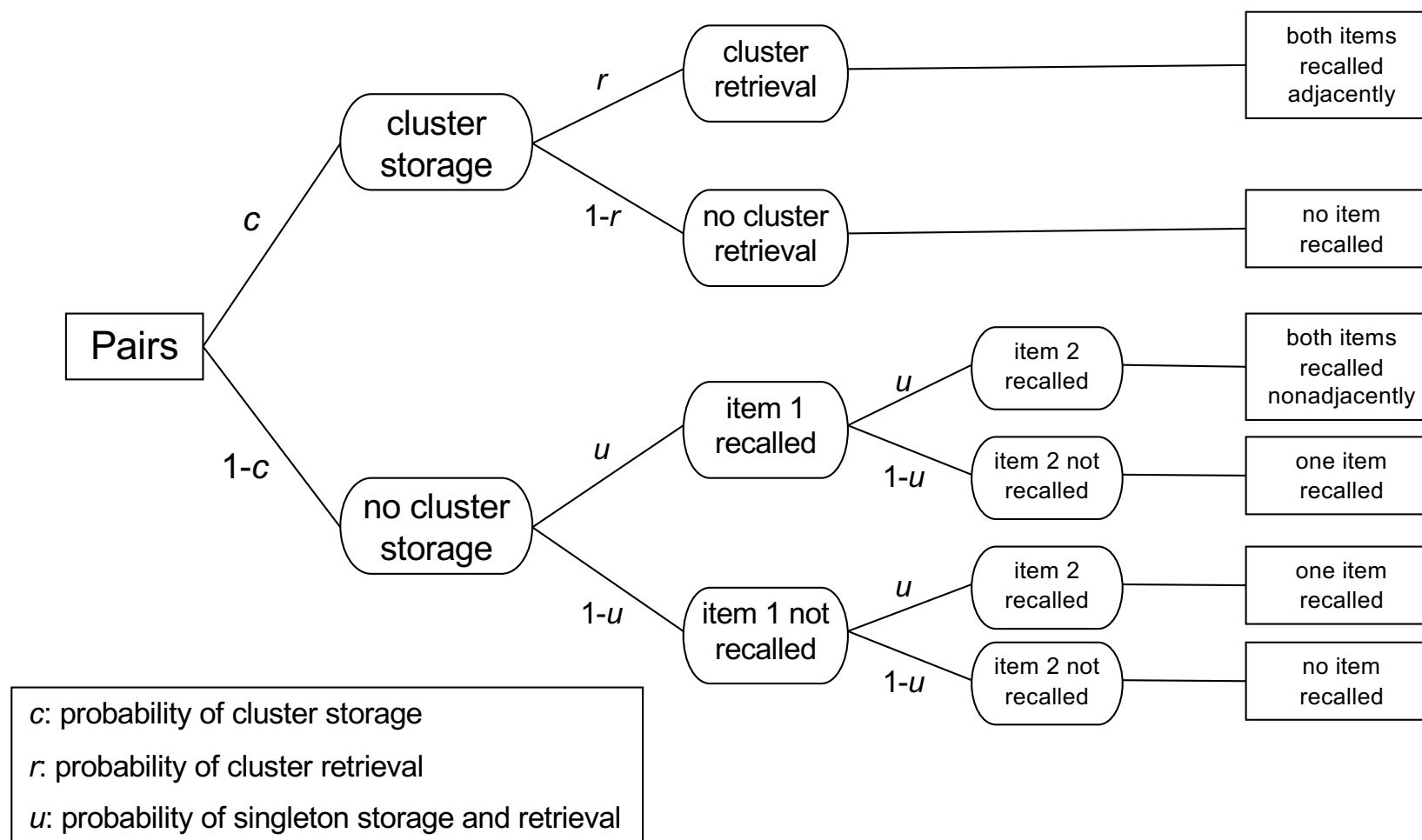
- *Empirical Paradigm:*
  - Free recall of word list consisting of
    - Word pairs (e.g. „chair“ und „table“)
    - Singletons (e.g., „rose“ and no other flower).
    - „Primacy–“ and „Recency–Buffer“.
- *Two distributions of responses:*
  - pairs
  - singletons

# Scoring of responses

- Observation categories
  - For word pairs:
    - E1 — both words recalled adjacently
    - E2 — both words recalled nonadjacently
    - E3 — one word recalled
    - E4 — no word recalled
  - For singletons:
    - F1 — Recalled
    - F2 — Not recalled

# Pair-Clustering Model

(Batchelder & Riefer, 1980, 1986)



# Model equations

- *Word pairs:*
- $p(E_1) = c \cdot r$
- $p(E_2) = (1 - c) \cdot u^2$
- $p(E_3) = (1 - c) \cdot 2 \cdot u \cdot (1 - u)$
- $p(E_4) = c \cdot (1 - r) + (1 - c) \cdot (1 - u)^2$
- *Singletons:*
- $p(F_1) = u$
- $p(F_2) = 1 - u$

## 1.3) Model development

- Preliminary summary:
- Select a paradigm (e.g., a task)
- Define the conditions of the paradigm
- Define the category system for each condition
- List relevant processes/parameters
- Construct theoretically reasonable processing branches („trees“) for each condition
- Derive corresponding model equations.
- General rules:
  - As simple as possible!!
  - Ignore unlikely events



# Recommendation

- When developing a new model, make notes of potentially problematic assumptions:
  - Cluster retrieval is an „all or nothing“ process;
  - All words are equally difficult;
  - Words of a pair that is not stored as a cluster are stored and retrieved in a stochastically independent fashion;
  - Words of a pair that is not stored as a cluster are never recalled adjacently.
- Design your experiment in a way such that these assumptions are likely to be true.
- If the model misfits the data, consider these issues again and try to revise the model.

## 1.4) Formal model structure

- Multinomial models
- Parameterized multinomial models
- Multinomial processing tree (MPT) models
- Processing-tree representation of MPT models

# Multinomial Models

a) *One condition (simple multinomial model)*

- One variable with  $J$  categories and sample frequencies  $n_1, n_2, \dots, n_j, \dots, n_J$ .
- $\boldsymbol{\pi} = (p_1, p_2, \dots, p_J)$  is the vector of category probabilities.
- Given independent sampling, the sample frequencies follow a multinomial distribution:

$$p_{N,\boldsymbol{\pi}}(n_1, n_2, \dots, n_J) = \frac{N!}{n_1! n_2! \dots n_J!} p_1^{n_1} p_2^{n_2} \dots p_J^{n_J}$$

*b) Several conditions (joint multinomial model)*

- In each condition  $k$  ( $k = 1, \dots, K$ ), one categorical variable with  $J(k)$  categories is observed.
- For each of the  $K$  conditions, a simple multinomial model holds.
- Given independence between conditions, the overall probability of the sample frequencies across conditions is

$$p = \prod_{k=1}^K p_{N(k), \pi(k)}(n_{1(k)}, n_{2(k)}, \dots, n_{J(k)})$$

# Parameterized Multinomial Models

- The category probabilities  $p_1, p_2$  etc. are rewritten as functions of latent parameters  $\theta_1, \theta_2, \dots, \theta_S$
- Thus, in case of a simple multinomial model we have
  - $p_1 = f_1(\theta_1, \theta_2, \dots, \theta_S)$
  - $p_2 = f_2(\theta_1, \theta_2, \dots, \theta_S)$
  - ....
  - $p_J = f_J(\theta_1, \theta_2, \dots, \theta_S)$
- These equations are called model equations.
- The set of possible values of  $S$  latent parameters is called “parameter space”  $\Omega$  of the model.

# Multinomial Processing Tree (MPT) Models

- MPT models form a subclass of parameterized multinomial models.
- Additional assumptions:
  - 1) Each  $\theta_s$  is in  $[0, 1]$

# Multinomial Processing Tree (MPT) Models

- MPT models form a subclass of parameterized multinomial models.
- Additional assumptions:
  - 1) Each  $\theta_s$  is in  $[0, 1]$
  - 2) Structure of the model equations (Hu & Batchelder, 1994):

$$p_j = \sum_{i=1}^{I(j)} c_{ij} \prod_{s=1}^S \theta_s^{a_{ijs}} \cdot (1 - \theta_s)^{b_{ijs}}, \quad \sum_{j=1}^J p_j = 1, \quad \theta_s \in [0, 1]$$

where  $s$ : Parameter index

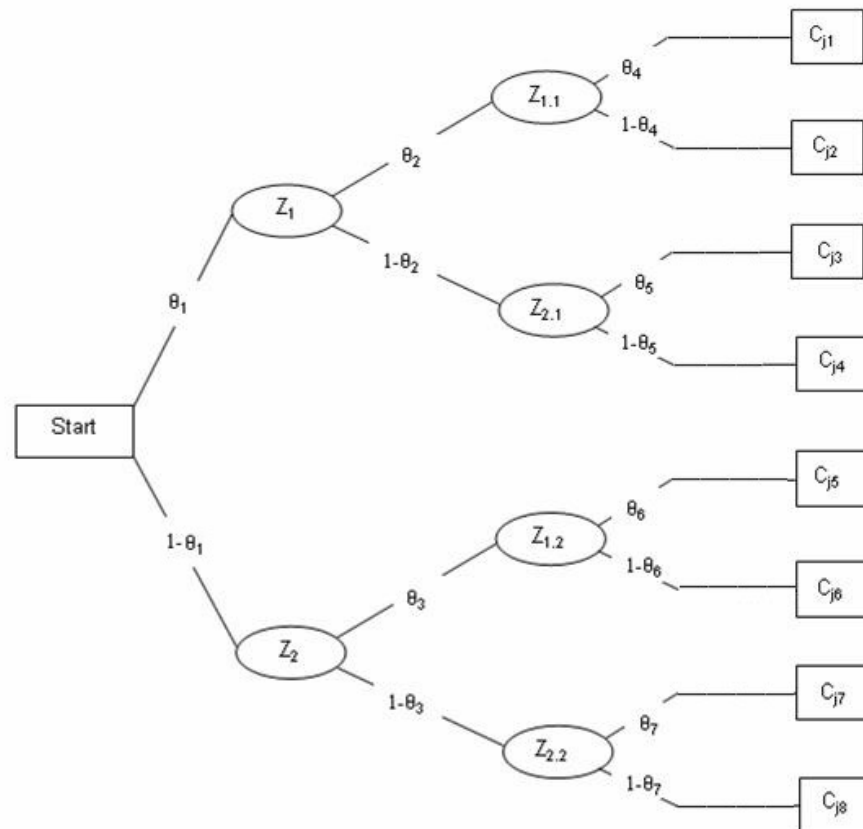
$j$ : Category index

$i$ : Branch index

$c_{ij}$ : positive real number

$a_{ijs}, b_{ijs}$ : nonnegative integer number (often 0 or 1)

Binary probabilistic event trees can always be translated in MPT-model equations:

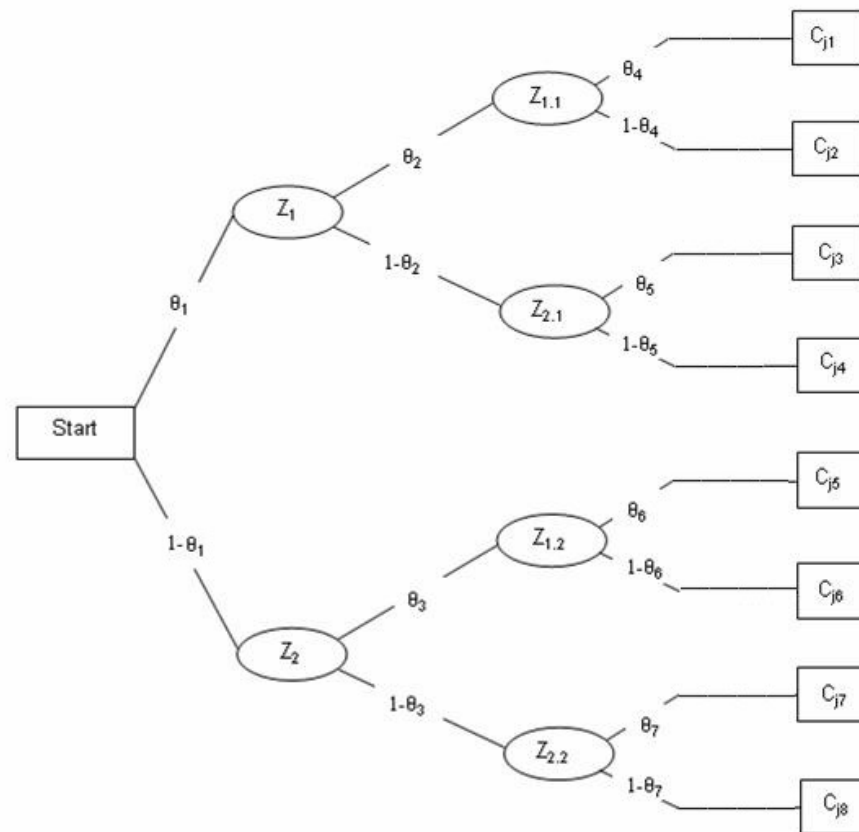




## ... as an aside ...

- Although any binary processing tree diagram uniquely determines a system of MPT model equations ...
- ... it is not true that any system of MPT model equations uniquely determines a specific processing tree diagram.
- Counter examples:
  - Level switching in independence models
  - Category switching given identical branch probabilities

Example: Assume  $\theta_1 = .5$ ,  $\theta_2 = \theta_3$ ,  $\theta_4 = \theta_6$



Another counter example is this  
MPT model:

$$p_1(\theta) = \theta^3$$

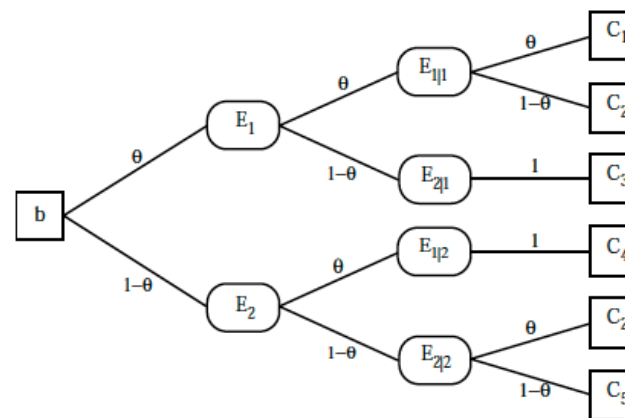
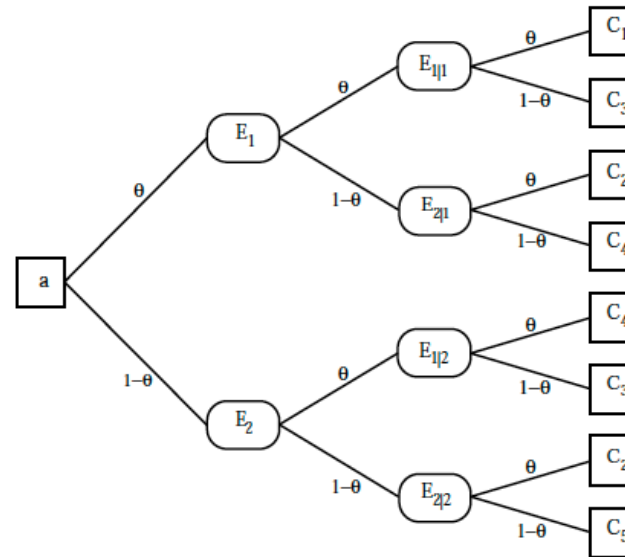
$$p_2(\theta) = \theta \cdot (1 - \theta)$$

$$p_3(\theta) = \theta \cdot (1 - \theta)$$

$$p_4(\theta) = \theta \cdot (1 - \theta)$$

$$p_5(\theta) = (1 - \theta)^3$$

# Two different processing trees compatible with the model on the previous slide:



# 1.5) Identifiability

Any MPT model equation system

$$p_j = \sum_{i=1}^{I(j)} c_{ij} \prod_{s=1}^S \theta_s^{a_{ijs}} \cdot (1 - \theta_s)^{b_{ijs}}, \quad \sum_{j=1}^J p_j = 1, \quad \theta_s \in [0, 1]$$

where  $s$  : Parameter index

$j$  : Category index

$i$  : Branch index

$c_{ij}$  : positive real number

$a_{ijs}, b_{ijs}$  : nonnegative integer number (often 0 or 1)

defines a mapping  $f: \Omega \rightarrow \mathbb{P}$

# Parameter Space and Data Space

$\Omega$  is called „Parameter Space“:

= Set of all possible parameter vectors

$P$  is called „Data Space“ (more precisely:  
space of category probabilities)

= Set of all possible category probability vectors

# Identifiability of an MPT model

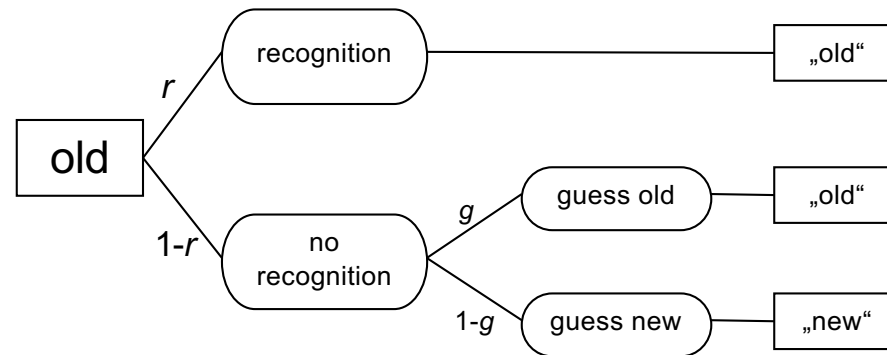
*Definition* (global identifiability):

An MPT model is globally identified if  $f$  is one-to-one.

*Definition* (local identifiability):

An MPT model is locally identified if  $f$  is one-to-one in the neighborhood of  $\theta_0$  in  $\Omega$ .

# One-High-Threshold recognition model (Blackwell, 1953)

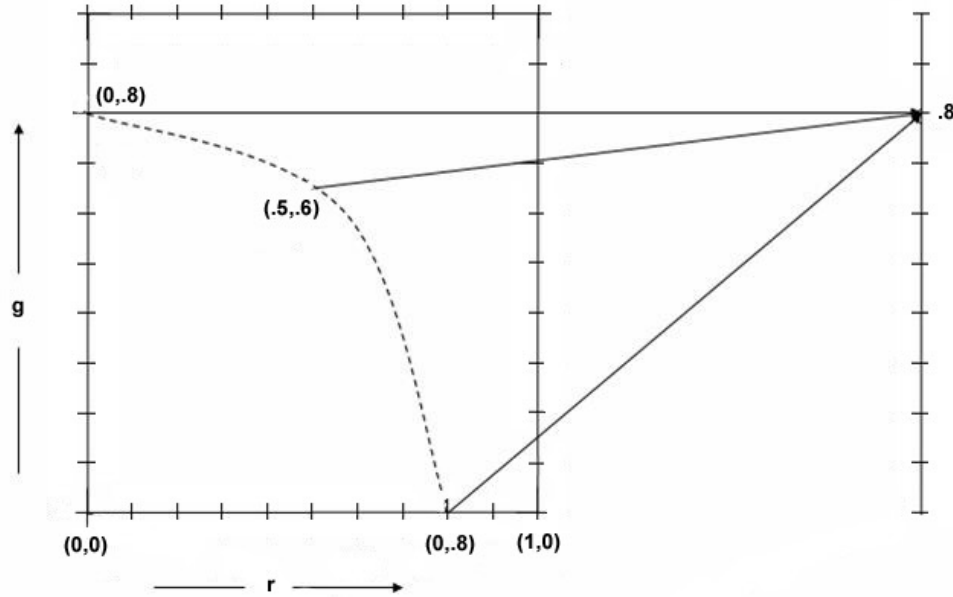


$r$ : probability of recognition

$g$ : probability of guessing old given recognition

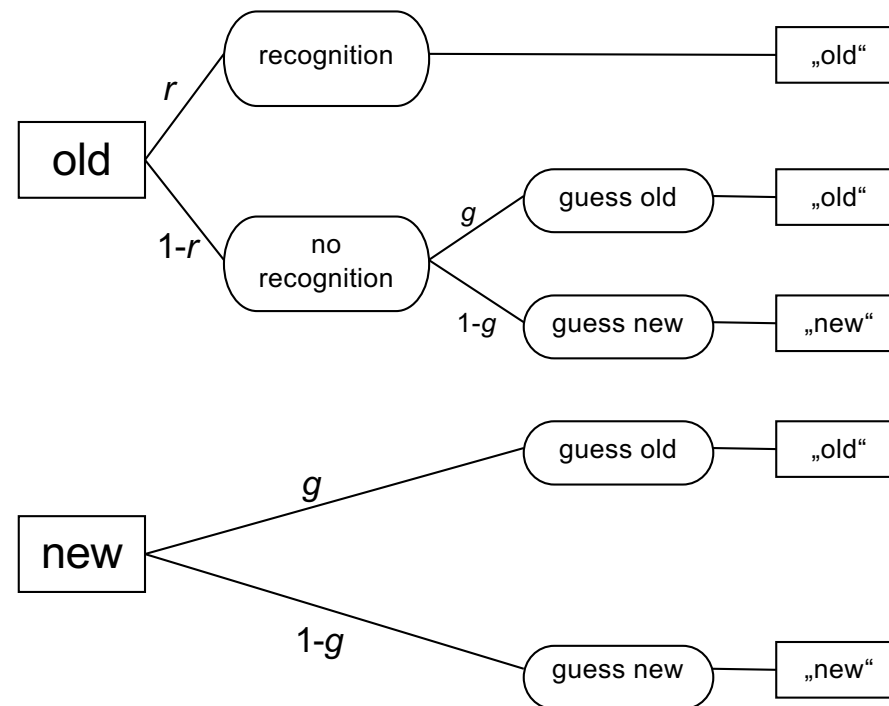


## Example 1 (nonidentifiability)



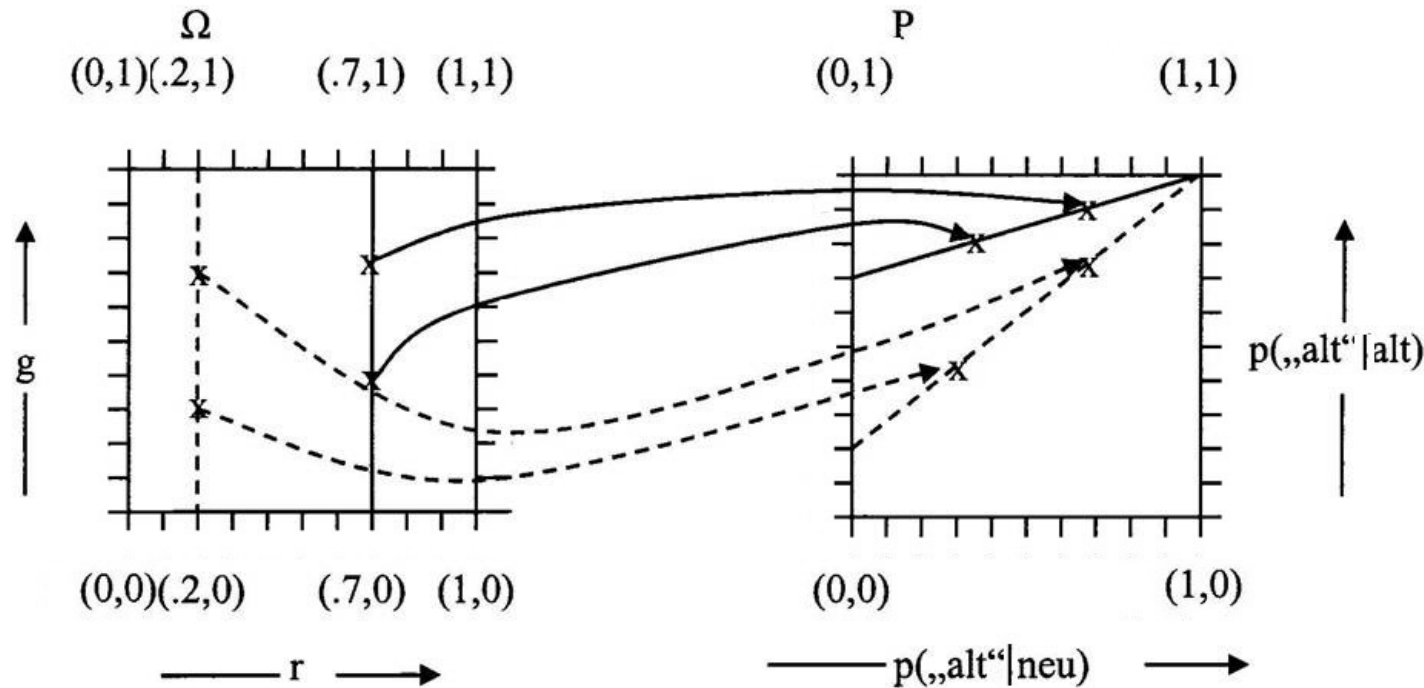
- One-high-Threshold-Modell limited to old items:
  - $p(\text{“alt”} \mid \text{old item}) = r + (1-r) \cdot g$
- Model is not identified.

# One-High-Threshold recognition model (Blackwell, 1953)



$r$ : probability of recognition  
 $g$ : probability of guessing old given recognition

# Example 2 (identifiability)



- One-high-Threshold-Model (old and new items)
  - $p(„alt“ | \text{old item}) = r + (1-r) \cdot g$
  - $p(„alt“ | \text{new item}) = g$
- Model is globally identified.

## Two important theorems:

- „*Observable branches*“:

A model is always globally identified if each of its branches terminates in a unique empirical category (Hu & Batchelder, 1994).

## Two important theorems:

- „*Observable branches*“:

A model is always globally identified if each of its branches terminates in a unique empirical category (Hu & Batchelder, 1994).

- „*No more parameters than degrees of freedom in the data*“:

A necessary but not sufficient condition of identifiability is

$$S \leq \sum_{k=1}^K (J(k) - 1).$$

# Jacobian Matrix

- Jacobian: Matrix of the 1st partial derivatives of the model equations with respect to parameters  $\theta_s$ ,  $s = 1, \dots, S$ .
- $r$ : maximum rank of the Jacobian across  $\Omega$
- If  $r < S$ , then the model is neither locally nor globally identified.
- If  $r = S$ , then the model is locally identified (but not necessarily globally).

# Remedies for nonidentifiable models

- Parameter constraints
  - Parameter fixations ( $\theta_s = c$ , with  $c = \text{constant}$ )
  - Equality constraints ( $\theta_s = \theta_{s'}$ )
- Increase the number of empirical categories
  - Add more conditions that require no (or few) additional parameters
  - Simple recipe that often works: Selective manipulations of parameters

# 1.6) Parameter Estimation

Find a parameter vector  $\theta = (\theta_1, \dots, \theta_s, \dots, \theta_S), \theta \in \Omega$ , such that the distance between the sample frequencies  $n_1, n_2, \dots, n_J$  and the expected category frequencies under the model,  $N \cdot p_1(\theta), N \cdot p_2(\theta), \dots, N \cdot p_J(\theta)$ , becomes a minimum.



# 1.6) Parameter Estimation

Find a parameter vector  $\theta = (\theta_1, \dots, \theta_s, \dots, \theta_S), \theta \in \Omega$ , such that the distance between the sample frequencies  $n_1, n_2, \dots, n_J$  and the expected category frequencies under the model,  $N \cdot p_1(\theta), N \cdot p_2(\theta), \dots, N \cdot p_J(\theta)$ , becomes a minimum.

Which distance measure?

# 1.6) Parameter Estimation

Find a parameter vector  $\theta = (\theta_1, \dots, \theta_s, \dots, \theta_S), \theta \in \Omega$ , such that the distance between the sample frequencies  $n_1, n_2, \dots, n_J$  and the expected category frequencies under the model,  $N \cdot p_1(\theta), N \cdot p_2(\theta), \dots, N \cdot p_J(\theta)$ , becomes a minimum.

Which distance measure?

The likelihood ratio statistic  $G^2$  and Pearson's  $\chi^2$  are used most often:

$$G^2(\theta) = 2 \sum_{j=1}^J n_j \ln \left( \frac{n_j}{N \cdot p_j(\theta)} \right) \quad \chi^2(\theta) = \sum_{j=1}^J \frac{[n_j - N \cdot p_j(\theta)]^2}{N \cdot p_j(\theta)}$$

Both distance measures are special cases of the Power-Divergence-family ( $PD_\lambda$ -Statistics) (Read & Cressie, 1988):

$$s_\lambda = \frac{2}{\lambda(\lambda + 1)} \sum_{k=1}^k \sum_{i=1}^{J(k)} n_{j(k)} \cdot \left[ \left( \frac{n_{j(k)}}{e_{j(k)}} \right)^\lambda - 1 \right]$$

Note that:

$$\begin{aligned} \text{Pearson-}\chi^2 &= S_{\lambda=1} \\ \text{Likelihood-ratio-}G^2 &= \lim_{\lambda \rightarrow 0} S_\lambda. \end{aligned}$$

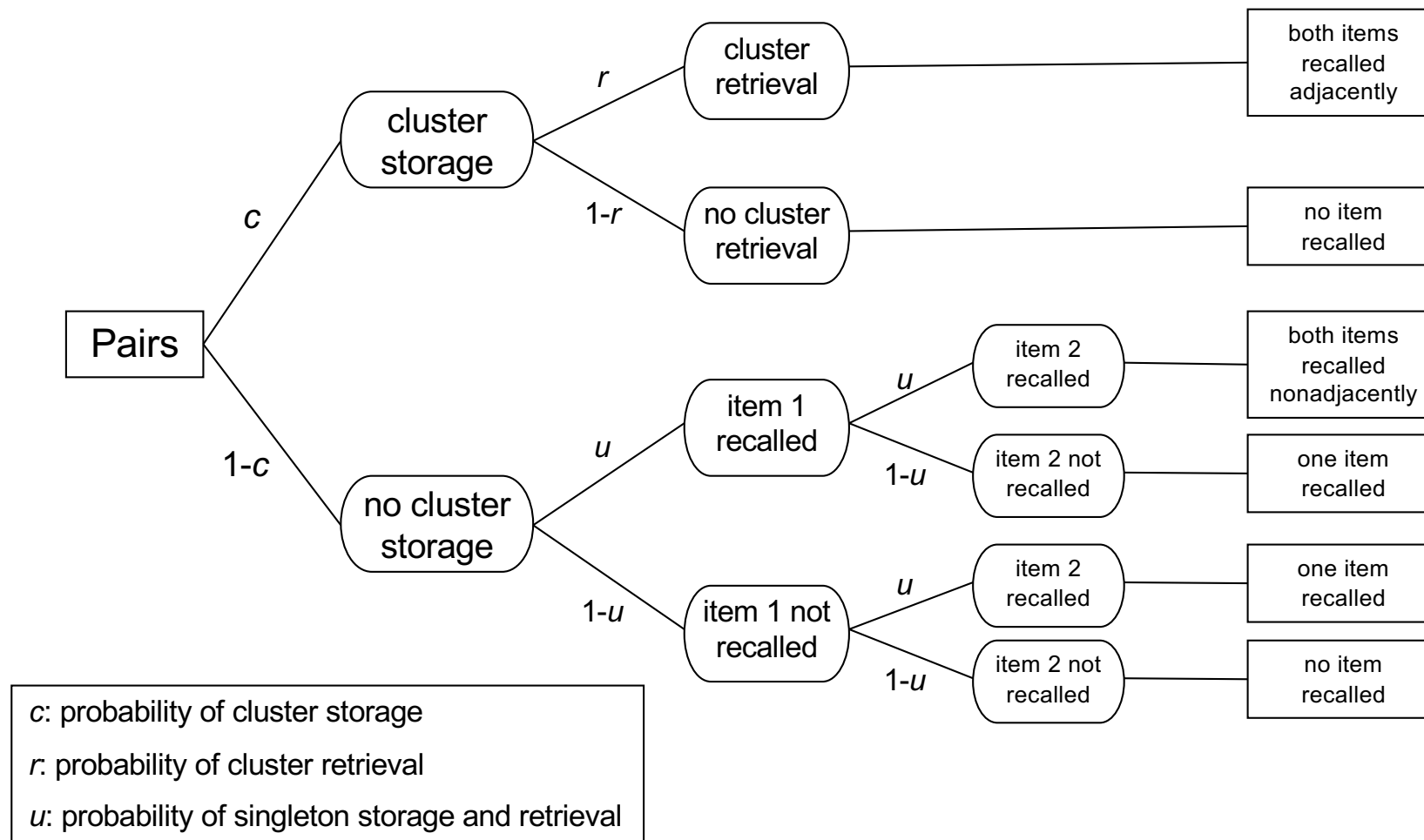
# What is the best estimation method?

- All  $S_\lambda$ -estimates have the same asymptotic properties:
  - Consistent, efficient, asymptotically unbiased
  - (multivariate) normal sampling distributions
- Choosing  $\lambda=0$  (i.e., minimizing  $G^2$ ) provides Maximum Likelihood (ML) estimates:
  - Sample frequencies have highest likelihood
  - Standard method in typical applications
- All  $S_\lambda$ -estimates are easily implemented using the robust Expectation-Maximization (EM) estimation algorithm.

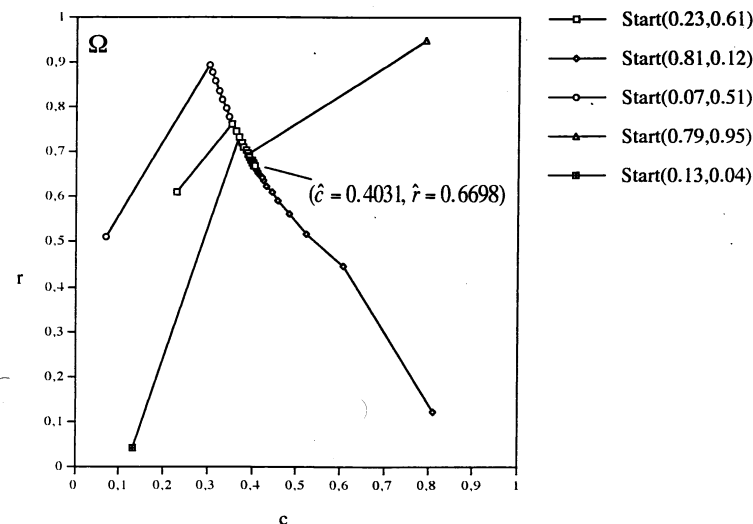
## *Expectation-Maximization-(EM) Algorithm*

- 1) Choose a (random) start vector  $\theta_i$ .
- 2) **E**xpectation)-Step:
  - Estimate the expected frequencies of the branches given  $\theta_i$  and the observed category frequencies  $n_{j(k)}$
- 3) **M**aximization)-Step:
  - Let  $i = i + 1$
  - Compute new  $S_\lambda$  estimates  $\theta_i$  given the expected frequencies from step 2)
- 4) Convergence ?
  - If  $\text{Abs}(\theta_i - \theta_{i-1}) > \text{Criterion}$  for some  $i$ , go back to step 2)
- 5) Otherwise accept  $\theta_i$  as final parameter estimates  $\hat{\theta}$

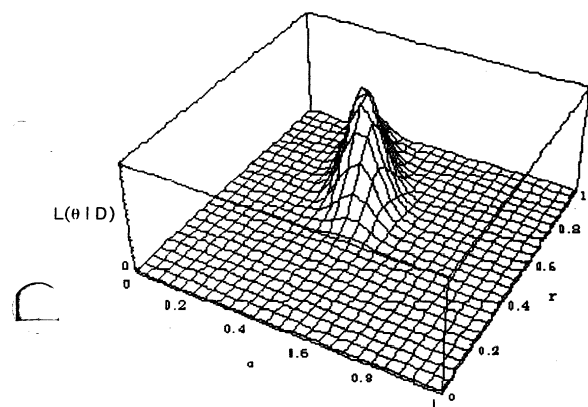
# Application of EM algorithm to Pair-Clustering Model



# Graphical Illustration of the EM-Algorithm applied to $c$ and $r$ in storage-retrieval model:



EM paths in  $\Omega$  for  $\lambda = 0$ , given five different starting values (simulated data).



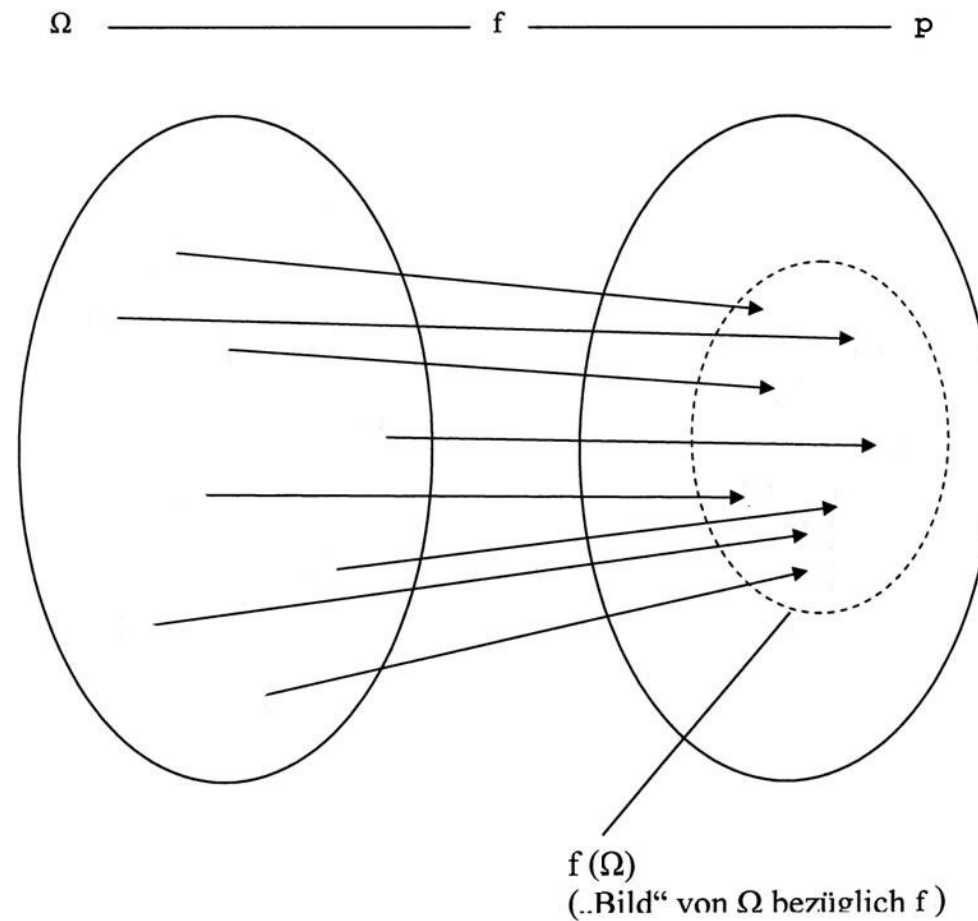
Likelihood (y-axis) as a function of  $c$  and  $r$  given the same simulated data as before.

# Bottom line

- Use  $G^2$ -minimization (= likelihood maximization) as the default estimation method.
- Repeat the estimation process several times using random start values to make sure that you really found a global maximum of the likelihood function.



# 1.7) Model assessment



How to test  $H_0: \pi \in f(\Omega)$  ?

Options:

$$\text{Pearson } \chi^2 = \sum_{k=1}^K \sum_{j=1}^{J(k)} (n_{j(k)} - e_{j(k)})^2 / e_{j(k)},$$

or

$$\text{Likelihood - ratio - } G^2 = 2 \cdot \sum_{k=1}^K \sum_{j=1}^{J(k)} n_{j(k)} \cdot \ln(n_{j(k)} / e_{j(k)});$$

$$\text{with } e_{j(k)} = p_{j(k)}(\hat{\theta}) \cdot N(k)$$

Both statistics are special cases of Read und Cressies (1988)  
 „Power Divergence“ – Statistics:

$$s_{\lambda} = \frac{2}{\lambda(\lambda + 1)} \sum_{k=1}^K \sum_{j=1}^{J(k)} n_{j(k)} * \left[ \left( \frac{n_{j(k)}}{e_{j(k)}} \right)^{\lambda} - 1 \right]$$

Again:

$$\text{Pearson-}\chi^2 = S_{\lambda=1}$$

$$\text{Likelihood-ratio-}G^2 = \lim_{\lambda \rightarrow 0} S_{\lambda}.$$

All  $S_{\lambda}$  – Statistics are asymptotically (i.e., for  $N \rightarrow \infty$ ) chi-square distributed under  $H_0$  (i.e., the model holds) with

$$\text{df} = \sum_{k=1}^K (J(k) - 1) - S.$$

# What is the best goodness-of-fit statistic?

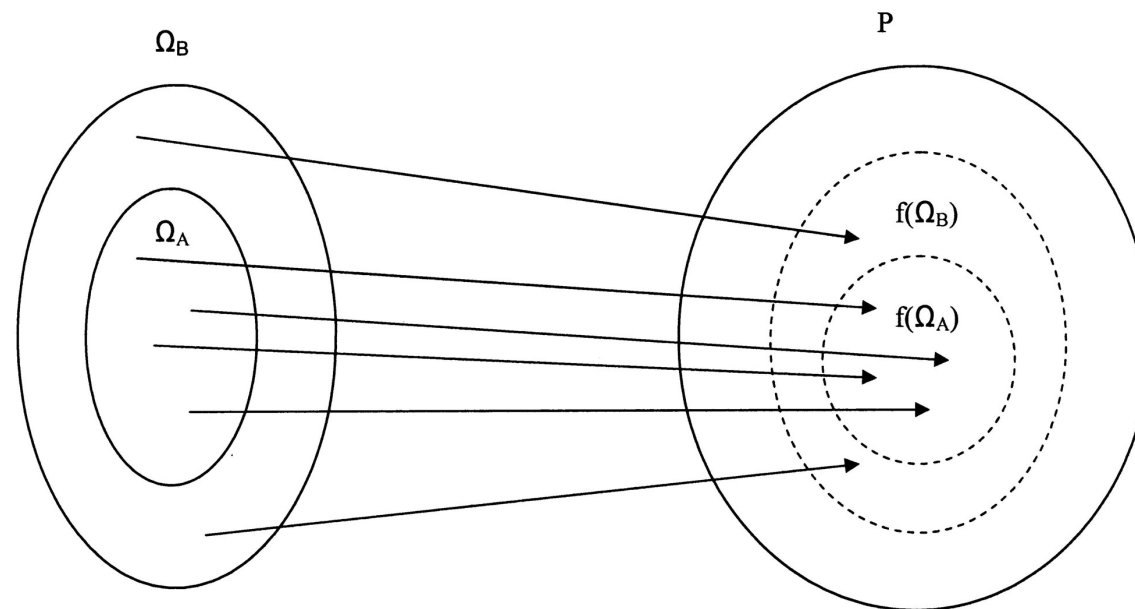
- In case of very small sample sizes,  $S_{\lambda=2/3}$  outperforms other  $S_{\lambda}$ -statistics (such as  $G^2 = S_{\lambda=0}$ ) in terms of accuracy of Chi-square approximation (cf. Read & Cressie, 1988).
- However, very small samples are typically less of a problem in MPT model applications (unless models are tested for single participants).
- Given the fact that  $G^2$  is a by-product of ML-parameter estimation,  $G^2$  can be recommended for moderate to large sample sizes.
- However, you cannot choose  $G^2$  in case of samples with zero cells.
- Remedies: 1) Ignore 0 cells, 2) add constant  $\varepsilon$  to all counts

# Model comparisons

- Assume that model B ( $M_B$ ) holds for your data:  $G^2(M_B)$  is not significant.
- Then often  $M_B$  is compared to an alternative model A ( $M_A$ ) that is a special case of  $M_B$ .
- $M_A$  is obtained by applying a parameter restriction to  $M_B$ , e.g., an equality constraint or a parameter fixation.
- How to decide whether  $M_A$  fits the data worse than  $M_B$ ?

Model  $M_A$  as a special case of  $M_B$ :

Illustration of model-specific mapping of  $\Omega_A$  and  $\Omega_B$  in  $P$ .



This is called a „hierarchical model family“.

# Model comparisons in hierarchical model families

- If both  $M_A$  and  $M_B$  hold then
  - $G^2_A$  is chi-square distributed with  $df_A$
  - $G^2_B$  is chi-square distributed with  $df_B$
  - $\Delta G^2_{A-B} := G^2_A - G^2_B$  is chi-square distributed with  $df_{A-B} = df_A - df_B$ .
- In other words,  $G^2$  is additive (irrespective of  $N$ ):
$$G^2_A = G^2_B + \Delta G^2_{A-B}$$
- This additivity property does not hold for other  $PD_\lambda$ -stats.
- Using  $\Delta G^2_{A-B}$ , it is easy to compare different models in a hierarchical model family using chi-square tests
- This is perhaps the strongest argument for relying on  $G^2$  for purposes of model fitting and testing.

# Model comparisons in nonhierarchical model families

- Unfortunately,  $\Delta G^2_{A-B} := G^2_A - G^2_B$  cannot be used in nonhierarchical model families.
- How to proceed then?



# Information-theoretic measures of goodness-of-fit

- Akaike Information Criterion (AIC):
  - $AIC(M_0) = -2 \cdot \ln(L(\theta; y)) + 2 \cdot S$
  - $\Delta AIC(M_0) = AIC(M_0) - AIC(\text{sat.}) = G^2(M_0) - 2 \cdot \text{df}(M_0)$
- Bayesian Information Criterion (BIC):
  - $BIC(M_0) = -2 \cdot \ln(L(\theta; y)) + S \cdot \ln(N)$
  - $\Delta BIC(M_0) = BIC(M_0) - BIC(\text{sat.}) = G^2(M_0) - \text{df}(M_0) \cdot \ln(N)$
- Rule:
  - Choose the model with the smaller AIC / BIC
- Recommendation:
  - Fit is fine if  $\Delta AIC$  or  $\Delta BIC < 0$ .

# 1.8) Selected literature

- Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, 6, 57-86.
- Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie-Journal of Psychology*, 217, 108-124.
- Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, 59, 21-47.
- Kuhlmann, B., Erdfelder, E., & Moshagen, M. (2019). Testing interactions in multinomial processing tree models. *Frontiers in Psychology*, 10, 2364
- Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, 42, 42-54.

- Klauer, K.C. (2006). Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika*, 71, 7–31.
- Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, 95, 318-339.
- Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods*, 45, 560–575.
- Schmidt, O., Erdfelder, E., & Heck, D. W. (2022, March 14). Tutorial on Multinomial Processing Tree Modeling: How to Develop, Test, and Extend MPT Models.  
<https://doi.org/10.31234/osf.io/gh8md>
- Stahl, C. & Klauer, K.C. (2007). HMMTree: A computer program for latent-class hierarchical multinomial processing-tree models. *Behavior Research Methods*.